Retrospective for the Dynamic Sensorium Competition for predicting large-scale mouse primary visual cortex activity from videos

Polina Turishcheva[*,1,✉], **Paul G. Fahey**[*,2–5,✉], **Michaela Vystrčilová**[1], **Laura Hansel**[1], **Rachel Froebe**[2–5], **Kayla Ponder**[2], **Yongrong Qiu**[1,3–5], **Konstantin F. Willeke**[1,6,7], **Mohammad Bashiri**[1,6,7] **Ruslan Baikulov**[8], **Yu Zhu**[9,10], **Lei Ma**[10], **Shan Yu**[9], **Tiejun Huang**[10], **Bryan M. Li**[11, 12], **Wolf De Wulf**[12], **Nina Kudryashova**[12], **Matthias H. Hennig**[12], **Nathalie L. Rochefort**[13,14], **Arno Onken**[12], **Eric Wang**[2], **Zhiwei Ding**[2], **Andreas S. Tolias**[2–5,15], **Fabian H. Sinz**[1,2,6,7], **Alexander S. Ecker**[1,16,✉]

[1] Institute of Computer Science and Campus Institute Data Science, University of Göttingen, Germany, [2] Department of Neuroscience & Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, Texas, USA, [3] Department of Ophthalmology, Byers Eye Institute, Stanford University School of Medicine, Stanford, CA, US, [4] Stanford Bio-X, Stanford University, Stanford, CA, US, [5] Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, US, [6] International Max Planck Research School for Intelligent Systems, Tübingen, Germany [7] Institute for Bioinformatics and Medical Informatics, Tübingen University, Germany, [8] lRomul, Russia, [9] Institute of Automation, Chinese Academy of Sciences, China [10] Beijing Academy of Artificial Intelligence, China [11] The Alan Turing Institute, UK [12] School of Informatics, University of Edinburgh, UK [13] Centre for Discovery Brain Sciences, University of Edinburgh, UK [14] Simons Initiative for the Developing Brain, University of Edinburgh, UK [15] Department of Electrical Engineering, Stanford University, Stanford, CA, US, [16] Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany.
[*] Equal contribution. [✉] turishcheva@cs.uni-goettingen.de, pgfahey@stanford.edu, ecker@cs.uni-goettingen.de

Abstract

Understanding how biological visual systems process information is challenging because of the nonlinear relationship between visual input and neuronal responses. Artificial neural networks allow computational neuroscientists to create predictive models that connect biological and machine vision. Machine learning has benefited tremendously from benchmarks that compare different models on the same task under standardized conditions. However, there was no standardized benchmark to identify state-of-the-art dynamic models of the mouse visual system. To address this gap, we established the SENSORIUM 2023 Benchmark Competition with dynamic input, featuring a new large-scale dataset from the primary visual cortex of ten mice. This dataset includes responses from 78,853 neurons to 2 hours of dynamic stimuli per neuron, together with behavioral measurements such as running speed, pupil dilation, and eye movements. The competition ranked models in two tracks based on predictive performance for neuronal responses on a held-out test set: one focusing on predicting in-domain natural stimuli and another on out-of-distribution (OOD) stimuli to assess model generalization. As part of the NeurIPS 2023 Competition Track, we received more than 160 model submissions from 22 teams. Several new architectures for predictive models were proposed, and the winning teams improved the previous state-of-the-art model by 50%. Access to the dataset as well as the benchmarking infrastructure will remain online at www.sensorium-competition.net.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-3777

1 Introduction

Understanding visual system processing has been a longstanding goal in neuroscience. One way to approach the problem are neural system identification approaches which make predictions of neuronal activity from stimuli or other sources quantitative and testable (reviewed in Wu et al., 2006). Various system identification methods have been used in systems neuroscience, including linear-nonlinear (LN) models (Simoncelli et al., 2004; Jones & Palmer, 1987; Heeger, 1992a,b), energy models (Adelson & Bergen, 1985), subunit models (Liu et al., 2017; Rust et al., 2005; Touryan et al., 2005; Vintch et al., 2015), Bayesian models (Walker et al., 2020; George & Hawkins, 2005; Wu et al., 2023; Bashiri et al., 2021), redundancy reduction models (Perrone & Liston, 2015), and predictive coding models (Marques et al., 2018). In recent years, deep learning models, especially convolutional neural networks (CNNs) trained on image recognition tasks (Yamins et al., 2014; Cadieu et al., 2014; Cadena et al., 2019; Pogoncheff et al., 2023) or predicting neural responses (Cadena et al., 2019; Antolík et al., 2016; Batty et al., 2017; McIntosh et al., 2016; Klindt et al., 2017; Kindel et al., 2019; Burg et al., 2021; Lurz et al., 2021; Bashiri et al., 2021; Zhang et al., 2018b; Cowley & Pillow, 2020; Ecker et al., 2018; Sinz et al., 2018; Walker et al., 2019; Franke et al., 2022; Wang et al., 2023; Fu et al., 2023; Ding et al., 2023), have significantly advanced predictive model performance. More recently, transformer-based architectures have emerged as a promising alternative (Li et al., 2023; Azabou et al., 2024; Antoniades et al., 2023).

In machine learning and beyond, standardized large-scale benchmarks foster continuous improvements in predictive models through fair and competitive comparisons (Dean et al., 2018). Within the realm of computational neuroscience, several benchmarks have been established recently. An early effort was the Berkeley Neural Prediction Challenge[1], which provided public training data and secret test set responses to evaluate models of neurons from primary visual cortex, primary auditory cortex and field L in the songbird brain. More recent efforts include Brain-Score (Schrimpf et al., 2018, 2020), Neural Latents '21 (Pei et al., 2021), Algonauts (Cichy et al., 2019, 2021; Gifford et al., 2023) and SENSORIUM 2022 (Willeke et al., 2022). However, except for the Berkeley Neural Prediction Challenge, which is limited to 12 neurons, no public benchmark existed that focused on predicting single neuron responses in the early visual system to video (spatio-temporal) stimuli.

Since we all live in a non-static world, dynamic stimuli are more relevant and our models should be able to predict neural responses over time in response to these time-varying inputs (Sinz et al., 2018; Wang et al., 2023; Batty et al., 2017; McIntosh et al., 2016; Zheng et al., 2021; Qiu et al., 2023; Hoefling et al., 2022; Vystrčilová et al., 2024). Even though recent high-throughput recording techniques have led to the release of large datasets like the MICrONS calcium imaging dataset (MICrONS Consortium et al., 2021) and Neuropixel datasets from the Allen Brain Observatory (de Vries et al., 2020; Siegle et al., 2021), the absence of a withheld test set and the corresponding benchmark infrastructure hinders a fair comparison between different models.

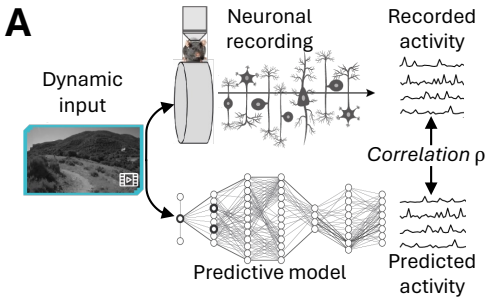
To fill this gap, we established the SENSORIUM 2023 competition, with the goal to compare large-scale models predicting single-neuron responses to dynamic stimuli. The NeurIPS 2023 competition received over 160 model submissions from 22 teams and resulted in new state-of-the-art predictive models that improved over the competition baseline by 50%. Moreover, these models also led to a 70% improved predictions on out-of-domain stimuli, suggesting that more predictive models on natural scenes also generalize better to other stimuli.

2 Sensorium Competition Overview

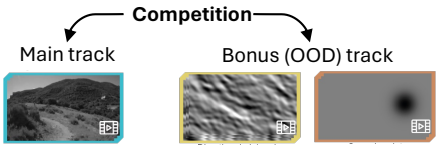
The goal of the competition was to advance models that predict neuronal responses of several thousand neurons in mouse primary visual cortex to natural and artificially generated movies. We collected and released a comprehensive dataset consisting of visual stimuli and corresponding neuronal responses for training (Section 3). This dataset included a dedicated test set, for which we released only the visual stimuli but withheld the neuronal responses to be able to compare models in a fair way (Fig. 1A). To assess model performance, participants submitted their predictions on the test set to our online benchmark website for evaluation and ranking against other submissions.[2] The

[1] https://neuralprediction.org/npc/con.php

[2] Our benchmark webpage is based on Codalab Competitions (Pavao et al., 2022) available under the Apache License 2.0 https://github.com/codalab/codalab-competitions



B



	Main track		Bonus track	
Model	Single-trial	Avg.	Single-trial	Avg.
DwiseNeuro	**0.291**	**0.542**	**0.221**	**0.429**
Dynamic-V1FM	0.265	0.493	0.183	0.336
ViV$_1$T	0.243	0.457	0.178	0.333
Ensemble baseline	0.197	0.371	0.129	0.241

C

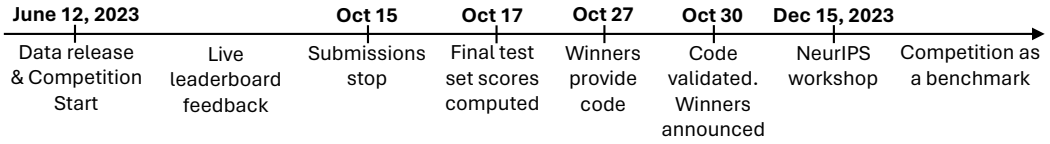


Figure 1: **A schematic illustration of the SENSORIUM competition. A:** Competition idea. We provide large-scale datasets of neuronal activity in the mice visual cortex in response to natural videos. The competition participants were tasked to find the best models to predict neuronal activity for a set of videos for which we withheld the ground truth. **B:** Tracks and leaderboard. **C:** Timeline.

test set consisted of two parts: a *live test set* and a *final test set*. The live test set was used during the competition to give feedback to the participants on their model's performance via a leaderboard. The final test set was used only after the end of submissions to determine the final winners (Fig. 1C). From each team only the best-performing submission on the live test set was evaluated on the final test set. To facilitate participation from both neuroscientists and machine learning practitioners, we developed user-friendly APIs that streamline data loading, model training, and submission.[3]

The competition consisted of two tracks: The *main track* and the *bonus track* (Fig. 1B). The main track entailed predicting responses on natural movie stimuli, the same type of stimuli available for model training, but different movie instances. The bonus track required predicting out-of-distribution (OOD) stimuli for which no ground truth responses of the neurons were provided in the training set. This bonus track tests a model's ability to generalize beyond the training data.

The competition ran from June 12 to October 15, 2023, culminating in a NeurIPS 2023 conference workshop where the winning teams presented their approaches and insights. The benchmark platform will continue to track advancements in developing models for the mouse primary visual cortex. In the following, we describe the dataset (Section 3) and evaluation metrics (Section 4), the baseline (Section 5) and winning models (Section 6) and report on the results and learnings (Section 7).

3 Dataset

We recorded[4] neuronal activity in response to natural movie stimuli as well as several behavioral variables, which are commonly used as a proxy of modulatory effects of neuronal responses (Niell & Stryker, 2010; Reimer et al., 2014). In general terms, neural predictive models capture neural responses $\mathbf{r} \in \mathbb{R}^{n \times t}$ of n neurons for t timepoints as a function $\mathbf{f}_\theta(\mathbf{x}, \mathbf{b})$ of both natural movie stimuli $\mathbf{x} \in \mathbb{R}^{w \times h \times t}$, where w and h are video width and height, and behavioral variables $\mathbf{b} \in \mathbb{R}^{k \times t}$, where $k = 4$ is the number of behavioral variables (see below).

Movie stimuli. We sampled natural dynamic stimuli from cinematic movies and the Sports-1M dataset (Karpathy et al., 2014), as described by MICrONS Consortium et al. (2021). Following earlier work (Wang et al., 2023), we showed five additional stimulus types for the bonus track (Fig. 2b): directional pink noise (MICrONS Consortium et al., 2021), flashing Gaussian dots, random dot kinematograms (Morrone et al., 2000), drifting Gabors (Petkov & Subramanian, 2007), and natural

[3] https://github.com/ecker-lab/sensorium_2023

[4] Full neuroscience methods are available in a technical report (Turishcheva et al., 2023).

https://doi.org/10.52202/079017-3777

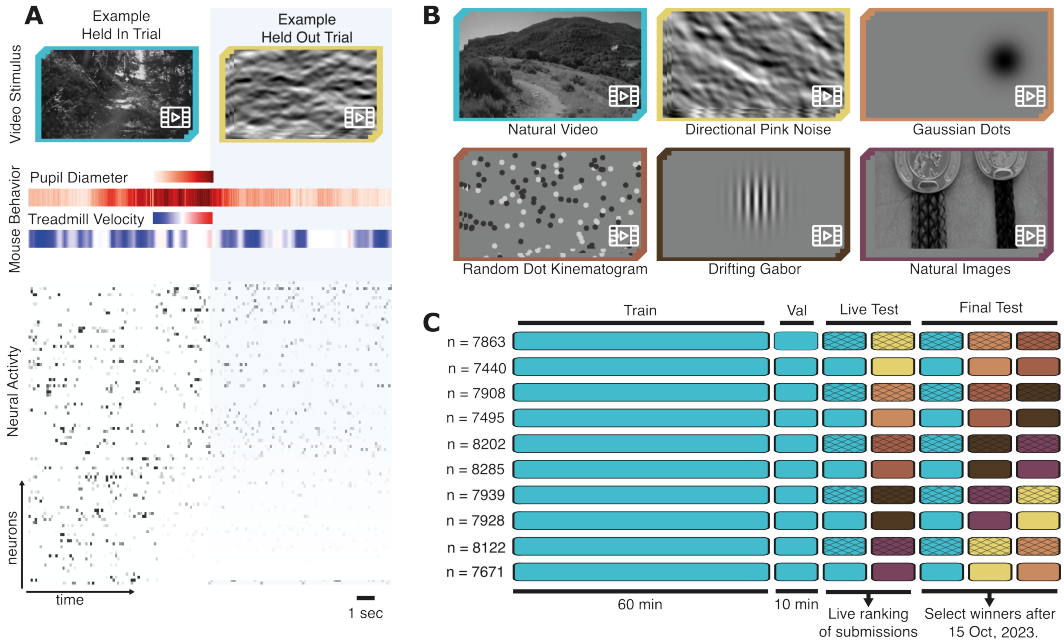


Figure 2: **Overview of the data. a**, Example stimulus frames, behavior (pupil position not depicted) and neural activity. **b**, Representative frames from natural video and five OOD stimuli. **c**, Stimulus composition (color) and availability for all five scans in ten animals. n is number of neurons per scan. The crossed elements were used for live and final test sets in the competition evaluation.

images from ImageNet (Russakovsky et al., 2015; Walker et al., 2019). Stimuli were converted to grayscale and presented to mice in clips lasting $\sim$ 8 to 11 seconds, at 30 frames per second.

Neuronal responses. Using a wide-field two-photon microscope (Sofroniew et al., 2016), we recorded the responses of excitatory neurons at 8 Hz in layers 2–5 of the right primary visual cortex in awake, head-fixed, behaving mice using calcium imaging. Neuronal activity was extracted as described previously (Wang et al., 2023) and upsampled to 30 Hz to be at the same frame rate as the visual stimuli. We also released the anatomical coordinates of the recorded neurons.

Behavioral variables. We measured four behavioral variables: *locomotion speed*, recorded from a cylindrical treadmill at 100 Hz and resampled to 30 Hz, and *pupil size, horizontal and vertical pupil center position*, each extracted at 20 Hz from video recordings of the eye and resampled to 30 Hz.

Datasets and splits. Our complete dataset consists of ten recordings from ten different animals, in total containing the activity of 78,853 neurons to a total of $\sim$1200 minutes of dynamic stimuli, with $\sim$120 minutes per recording. Five out of the ten recordings contributed to the live and final test sets. The recordings were collected and released explicitly for this competition. None of them had been published before. Each recording had four components (Fig. 2c):

Training set: 60 minutes of natural movies, one repeat each (60 minutes total).

Validation set: 1 minute of natural movies, ten repeats each (10 minutes total).

Live test set: 1 minute of natural movies and 1 minute of OOD stimuli, ten repeats each (20 minutes total). Each OOD stimulus type is presented only in one of the five recordings.

Final test set: 1 minute of natural movies and 2 minutes of OOD stimuli, ten repeats each (30 minutes total). Each OOD stimulus type is presented in two of the five recordings.

For the training set and validation set, the stimulus frames, neuronal responses, and behavioral variables are released for model training and evaluation by the participants, and are not included in the competition performance metrics. For the five mice included in the competition evaluation, the train and validation sets contain only natural movies but not the OOD stimuli. For the other five mice, all stimuli and responses, including test sets and OOD stimuli, were released.

Code and data availability. The competition website and data are available at https://www.sensorium-competition.net/. Starter kit and benchmark code are available at https://github.com/ecker-lab/sensorium_2023.

4 Competition evaluation

Similar to SENSORIUM 2022, we used the correlation coefficient between predicted and measured responses to evaluate the models. Since it is bounded between -1 and 1, the correlation coefficient is straightforward to interpret. Because neuronal responses fluctuate from trial to trial, the correlation between model predictions and single-trial neuronal responses typically do not reach the upper bound of 1 even for a perfect model. This trial-to-trial variability can be reduced by averaging over repeated presentations of the same stimulus. However, in this case, also the contributions from behavioral states are reduced since these cannot be repeated easily during uncontrolled behavior. We therefore computed two metrics: *single-trial correlation* and *correlation to average*.

Single-trial correlation, ρ_{st}, on the natural video final test set was used to determine competition winners for the main track. We also computed the single-trial correlation metric for each of the five OOD stimulus types in the test sets separately. The average single-trial correlation across all five final OOD test sets was used to determine the competition winner for the bonus track. Single trial correlation is sensitive to variation between individual trials and computes correlation between single-trial model output (prediction) o_{ij} and single-trial neuronal responses r_{ij}, as

$$\rho_{\text{st}} = \text{corr}(\mathbf{r}_{\text{st}}, \mathbf{o}_{\text{st}}) = \frac{\sum_{i,j}(r_{ij} - \bar{r})(o_{ij} - \bar{o})}{\sqrt{\sum_{i,j}(r_{ij} - \bar{r})^2 \sum_{i,j}(o_{ij} - \bar{o})^2}}, \tag{1}$$

where r_{ij} is the i-th frame of j-th video repeat, o_{ij} is the corresponding prediction, which can vary between stimulus repeats as the behavioral variables are not controlled. The variable $\bar{r}$ is the average response to all the videos in the corresponding test subset across all stimuli and repeats, and $\bar{o}$ is the average prediction for the same videos and repeats. The single-trial correlation ρ_{st} was computed independently per neuron and then averaged across all neurons to produce the final metric.

Correlation to average, ρ_{ta}, provides a more interpretable metric by accounting for trial-to-trial variability through averaging neuronal responses over repeated presentations of the same stimulus. As a result, a perfect model would have a correlation close to 1 (not exactly 1, since the average does not completely remove all trial-to-trial variability). However, correlation to average does not measure how well a model accounts for stimulus-independent variability caused by behavioral fluctuations.

We calculate ρ_{ta} in the same way as ρ_{st}, but we first average the responses and predictions per frame across all video repeats, where $\bar{r}_i$ is a response averaged over stimulus repeats for a fixed neuron:

$$\rho_{\text{ta}} = \text{corr}(\mathbf{r}_{\text{ta}}, \mathbf{o}_{\text{ta}}) = \frac{\sum_{i}(\bar{r}_i - \bar{r})(\bar{o}_i - \bar{o})}{\sqrt{\sum_{i}(\bar{r}_i - \bar{r})^2 \sum_{i}(\bar{o}_i - \bar{o})^2}}, \tag{2}$$

The initial 50 frames of predictions and neuronal responses were excluded from all metrics calculations. This allowed a "burn-in" period for models relying on history to achieve better performance.

5 Baseline models

SENSORIUM 2023 was accompanied by three model baselines, representing the state of the art in the field at the beginning of the competition:

GRU baseline is a dynamic model with a 2D CNN core and gated recurrent unit (GRU) inspired by earlier work (Sinz et al., 2018), but with more recently developed Gaussian readouts (Lurz et al., 2021), which improves performance. Conceptually, the 2D core transforms each individual frame of the video stimulus into a feature space which is subsequently processed by a convolutional GRU across time. The Gaussian readout then learns the spatial preference of each neuron in the visual space ("receptive field"), by learning the position at which a vector from the feature space is extracted by bilinear interpolation from the four surrounding feature map locations. This latent vector is multiplied with a weight vector learned per neuron ("embedding") and put through a rectifying nonlinearity to predict the activity of the neuron at each time step.

https://doi.org/10.52202/079017-3777

Model	Main track		Bonus track	
	single-trial ρ_{st} ↑	average ρ_{ta} ↑	single-trial ρ_{st} ↑	average ρ_{ta} ↑
DwiseNeuro	**0.291**	**0.542**	**0.221**	**0.429**
Dynamic-V1FM	0.265	0.493	0.183	0.336
ViV1T	0.243	0.457	0.178	0.333
Ensemble baseline	0.197	0.371	0.129	0.241
Factorized baseline	0.164	0.321	0.121	0.223
GRU baseline	0.106	0.207	0.059	0.106

Table 1: Model performance of competition winners and baselines on both tracks.

Factorized baseline is a dynamic model (Vystrčilová et al., 2024; Hoefling et al., 2022) with a 3D factorized convolution core and Gaussian readouts. In contrast to the GRU baseline, where the 2D CNN core does not interact with the temporal component, the factorized core learns both spatial and temporal filters in each layer.

Ensembled baseline is an ensembled version of the above factorized baseline over 14 models. Ensembling is a well-known tool to improve the model performance in benchmark competitions (Allen-Zhu & Li, 2023). As we wanted to encourage participants to focus on novel architectures and training methods beyond simple ensembling, only entries outperforming the ensembled baseline were considered candidates for competition winners.

Training. All baseline models were trained with batch size 40 in the following way: For each of the 5 animals 8 video snippets consisting of 80 consecutive video frames starting at a random location within the video were passed and the gradient accumulated over all animals before performing optimizing step. We used early stopping with a patience of 5.

6 Results and Participation

In the four-month submission period, out of 44 registered teams, 22 teams submitted a combined total of 163 models (main track: 22 teams, 134 submissions, bonus track: 5 teams and 29 submissions). The strong baseline models were surpassed in both tracks by 48% and 70%, respectively (Table 1). Notably, the winning model – DwiseNeuro – outperformed all other models on both tracks by a fairly decent margin, and the difference seemed even stronger on the out-of-domain data than on the main track. In contrast, the runner-up solution – Dynamic-V1FM – had somewhat of an edge over the third place – ViV1T – on the main track, but both were on par on the out-of-domain data (Table 1). In the following we describe the three winning teams' approaches.

6.1 First place: DwiseNeuro

Architecture. DwiseNeuro consists of three main parts: core, cortex, and readouts. The core consumes sequences of video frames and mouse behavior activity in separate channels, processing temporal and spatial features. Produced features are aggregated with global average pooling over space. The cortex processes the pooled features independently for each timestep, increasing the channels. Finally, each readout predicts the activation of neurons for the corresponding mouse.

Core. The first layer of the module is the stem. It is a point-wise 3D convolution for increasing the number of channels, followed by batch normalization. The rest of the core consists of factorised inverted residual blocks (Tan & Le, 2019; Sandler et al., 2018) with a `narrow -> wide -> narrow` channel structure (Fig. 3A). Each block uses (1) absolute positional encoding (Vaswani et al., 2017) to compensate for spatial pooling after the core, (2) factorized (1+1) convolutions (Tran et al., 2018), (3) parameter-free shortcut connections interpolating spatial sizes and repeating channels if needed, (4) squeeze-and-excitation mechanism (Hu et al., 2018) to dynamically recalibrate channel-wise features, (5) DropPath regularization (Larsson et al., 2016; Huang et al., 2016) that randomly drops the block's main path for each sample in the batch. Batch normalization is applied after each layer. SiLU activation (Elfwing et al., 2018) is used after expansion and depth-wise convolutions.

Cortex. Spatial information accumulated through positional encoding was compressed by spatial global average pooling after the core, while the time dimension was unchanged. The idea of the

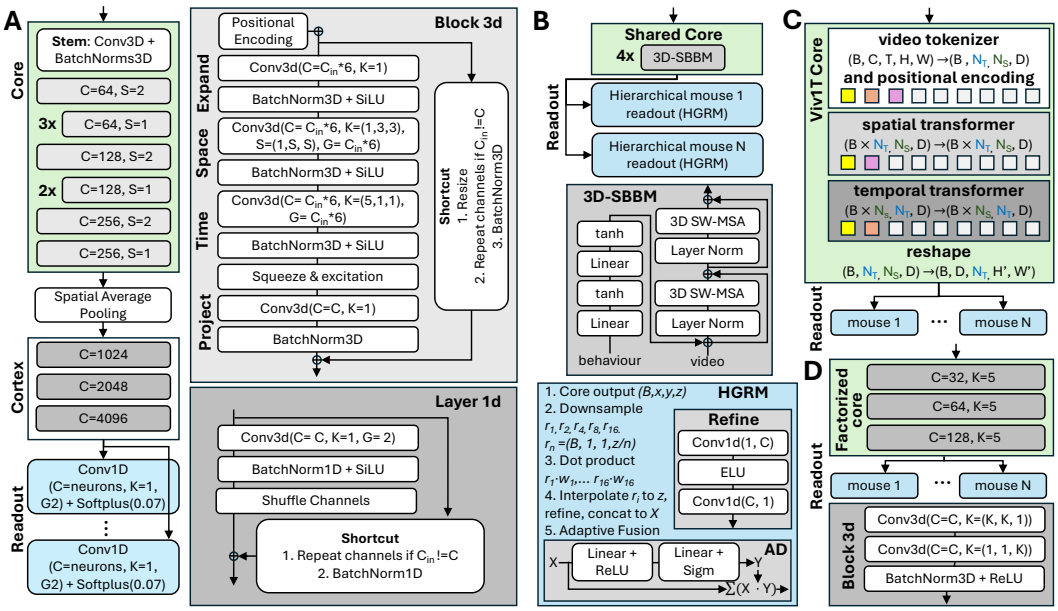


Figure 3: **Architectures of winning solutions.** Across all subplots: C: number of output channels in convolution layers, C_{in}: number of input channels, K: size of convolution kernels, S: stride, G: number of groups for convolution channels, B: batch size. Core: green, readout: blue. **A:** DwiseNeuro. The core is based on 3D factorised convolutions. The only solution whose readout was not based on the Gaussian readout (Lurz et al., 2021). **B:** Dynamic-V1FM. The core is transformer-based, the Gaussian readout is extended to look in different resolution to the core output, then to fuse different resolutions. Here w represents the readout linear weights learnt for each neuron. **C:** ViV1T. The idea is to replace the core with a spatiotemporal transformer. **D:** Ensembled factorized baseline.

"cortex" is to smoothly increase the number of channels before the readout and there is no exchange of information across time. First, the channels are split into two groups, then each group's channels are doubled as in a fully connected layer. Next, the channels are shuffled across the groups and concatenated. The implementation uses 1D convolution with two groups and kernel size one, with shuffling as in Zhang et al. (2018a). This procedure is repeated three times. Batch normalization, SiLU activation, and shortcut connections with stochastic depth were applied similarly to the core.

Readout. The readout is independent for each session, represented as a single 1D convolution with two groups and kernel size 1, 4096 input channels and the number of output channels equal to the number of neurons per mouse. It is followed by softplus activation as in Hoefling et al. (2022).

Training. The main changes compared to the baseline are introducing CutMix data augmentation (Yun et al., 2019), removing normalization, and padding the frames to 64×64 pixels. For more details on the training recipe, see Appendix A.1.

Code. Code is available at `https://github.com/lRomul/sensorium`

6.2 Second place: Dynamic-V1FM

Architecture. Dynamic-V1FM (Dynamic V1 Functional Model), follows the pipeline proposed by Wang et al. (2023). It incorporates a *shared core* module across mice and an *unshared readout* module for individual mice. The shared core module comprises four blocks of Layer Norms and 3D window based multi-head self-attention (MSA), inspired by the 3D swin transformer block (Liu et al., 2022) combined with a two-layer behavioral multi-layer perceptron (MLP) module (Li et al., 2023). The readout module is a Hierarchical Gaussian Readout Module (HGRM), which extends the Gaussian readout module (Lurz et al., 2021) by introducing a multi-layer design before the final linear readout (Fig. 3B).

Ensemble Strategy. As the readout module could support up to five levels of features and original layer is not downsampled and is always used as a base, four combinations of low-resolution features were traversed, resulting in $C_4^4+C_4^3+C_4^2+C_4^1 = 1+4+6+4 = 15$ models, where C_n^k is a binomial

https://doi.org/10.52202/079017-3777

Model	GPU	GPU memory	Batch Size	Wall Time
DwiseNeuro	2 × RTX A6000	48 Gb	32	12h
Dynamic-V1FM	8 × 2080Ti GPU	11 Gb	32	24h
ViV1T	1 × Nvidia A100	40 Gb	60	20h
Factorized baseline	1 × RTX A5000	24 Gb	40	8h
GRU baseline	1 × RTX A5000	24 Gb	40	10h

Table 2: Training time for a single model (before ensembling).

coefficient $C_n^k = \frac{n!}{k!(n-k)!}$ with n elements and k combinations. Feature enhancement modules were also added to the low-resolution part of these 15 models, but the performance improvement was insignificant. As another set of 15 candidate models, they were included in the subsequent average ensemble strategy. A model with the original Gaussian readout module was also trained as a baseline. The aforementioned 31 models were trained with a fixed random seed of 42, followed by an average ensemble of their predictions. For the final results of both competition tracks (the main track and the out-of-distribution track), the same model and ensemble strategy were used.

Code. Code is available at `https://github.com/zhuyu-cs/Dynamic-VFM`.

6.3 Third place: ViV1T

Architecture. The Vision Transformer (ViT, Dosovitskiy et al. 2021) was shown to be competitive in predicting mouse V1 responses to static stimuli (Li et al., 2023). Here, a factorized Transformer (ViV1T) core architecture was proposed, based on the Video Vision Transformer by Arnab et al. (2021), to learn a shared visual representation of dynamic stimuli across animals. The ViV1T core contains three main components: (1) a tokenizer that concatenates the video and behaviour variables over the channel dimensions and extracts overlapping tubelet patches along the temporal and spatial dimensions, followed by a factorized positional embedding which learns the spatiotemporal location of each patch; (2) a spatial Transformer which receives the tubelet embeddings and learns the spatial relationship over the patches within each frame; (3) a temporal Transformer receives the spatial embedding and learns a joint spatiotemporal representation of the video and behavioural information. This factorized approach allows to apply the self-attention mechanism over the space and time dimensions separately, reducing the size of the attention matrix and, hence, compute and memory costs. Moreover, the vanilla multi-head attention mechanism is replaced by FlashAttention-2 (Dao, 2023) and parallel attention (Wang, 2021) to further improve model throughput.

Training. The model was trained on all labeled data from the five original mice and the five competition mice. To isolate the performance differences solely due to the core architecture, the same shifter module, Gaussian readout (Lurz et al., 2021), data preprocessing and training procedure as the factorized baseline were employed. Finally, a Bayesian hyperparameter search (Akiba et al., 2019) of 20 iterations was performed to find an optimised setting for the core module (see Table 3).

Ensemble. The final submission was an average output of 5 models, initialized with different seeds.

Code. Code is available at `https://github.com/bryanlimy/ViV1T`.

7 Discussion

Different competition submissions explored different architectures. All winners employed architectures distinct from the baseline, but stayed roughly within the core-readout framework (Antolík et al., 2016; Klindt et al., 2017). Successful strategies included:

- Two out of three winning teams utilized transformer-based cores.
- Two teams also modified the readouts, but no team explicitly modeled temporal processing or interaction between neurons in the readout.
- However, the "cortex" module of the winning solution introduced several layers of nonlinear processing after spatial average pooling, effectively allowing all-to-all interactions.
- The winning solution kept the factorized 3D convolutions while introducing methods from computer vision models, such as skip connections and squeeze-and-excitation blocks.

These observations suggest that classic performance-boosting methods from computer vision are also helpful to boost the performance for neural predictive models. However, the impact of such architectural changes on the biologically meaningful insights, such as in Franke et al. (2022); Burg et al. (2021); Ustyuzhaninov et al. (2022), still needs to be validated and requires additional research.

Another observation is that all three winning solutions included a mechanism for all-to-all interaction: the winning solution in the "cortex", the other two by using a transformer-based core. Thus, although the CNN has originally been modeled after primary visual cortex (Fukushima, 1980), it does not seem to provide the best inductive bias for modeling, at least mouse V1. Long-range interactions appear to be important. The current data does not allow us to resolve whether these long-range interactions actually represent visual information, as expected from lateral connections within V1 (Gilbert & Wiesel, 1983, 1989), or from more global signals related to the animal's behavior (which is also fed as input to the core). This will be an interesting avenue for future research.

Moving from static images to dynamic inputs in SENSORIUM 2023 increased the participation threshold markedly because of the higher demands on compute and memory. As a result, many models cannot be trained on freely available resources such as Colab or Kaggle anymore (Table 2).

8 Conclusion

Predictive models are an important tool for neuroscience research and can deliver important insights to understand computation in the brain (Doerig et al., 2023). We have seen that systematically benchmarking such models on shared datasets can boost their performance significantly. With the SENSORIUM benchmark we have successfully established such an endeavor for the mouse visual system. The 2023 edition successfully integrated lessons from 2022, such as including an ensemble to encourage participants to focus on new architectures. However, there are still ways to go to achieve a comprehensive benchmark for models of the visual system. Future iterations could include, among others, the following aspects:

- Use chromatic stimuli in the mouse vision spectrum (Hoefling et al., 2022; Franke et al., 2022).
- Establish a benchmark for combining different data collection protocols (Azabou et al., 2024) or modalities (Antoniades et al., 2023).
- Focus not only on the predictive performance on natural scenes, but also on preserving biologically meaningful functional properties of neurons (Walker et al., 2019; Ustyuzhaninov et al., 2022).
- Extend beyond the primary visual cortex.
- Include more comprehensive measurements of behavioral variables.
- Include active behaviors of the animals.

We invite the research community to join us in this effort by continuing to participate in the benchmark and contribute to future editions.

Acknowledgments and Disclosure of Funding

The authors thank GWDG for the technical support and infrastructure provided. Computing time was made available on the high-performance computers HLRN-IV at GWDG at the NHR Center NHR@Göttingen. The project received funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via Project-ID 454648639 (SFB 1528), Project-ID 432680300 (SFB 1456), Project-ID 276693517 (SFB 1233) and Project-ID 390727645 (Cluster of Excellence Machine Learning New Perspectives for Science, EXC 2064/1); the European Research Council (ERC) under the European Unions Horizon Europe research and innovation programme (Grant agreement No. 101041669); the German Federal Ministry of Education and Research (BMBF) via the Collaborative Research in Computational Neuroscience (CRCNS) (FKZ 01GQ2107); National Institutes of Health (NIH) via National Eye Insitute (NEI) grant RO1-EY026927, NEI grant T32-EY002520; National Institute of Mental Health (NIMH) and National Institute of Neurological Disorders and Stroke (NINDS) grant U19-MH114830, NINDS grant U01-NS113294, and NIMH grants RF1-MH126883 and RF1-MH130416; National Science Foundation (NSF) NeuroNex grant 1707400; Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract no. D16PC00003; Defense Advanced Research Projects Agency (DARPA), Contract No. N66001-19-C-4020.

 https://doi.org/10.52202/079017-3777

The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The content is solely the responsibility of the authors and does not necessarily represent the official views of any funding agency.

FHS is supported by the Carl-Zeiss-Stiftung and an AWS Machine Learning research award. MFB and KFW were supported by the International Max Planck Research School for Intelligent Systems. BML and WDW were supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am.*, *2*(2), 284–299.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Allen-Zhu, Z., & Li, Y. (2023). Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv*.

Antolík, J., Hofer, S. B., Bednar, J. A., & Mrsic-Flogel, T. D. (2016). Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLoS computational biology*, *12*(6), e1004927.

Antoniades, A., Yu, Y., Canzano, J., Wang, W., & Smith, S. L. (2023). Neuroformer: Multimodal and multitask generative pretraining for brain data. *arXiv preprint arXiv:2311.00136*.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, (pp. 6836–6846).

Azabou, M., Arora, V., Ganesh, V., Mao, X., Nachimuthu, S., Mendelson, M., Richards, B., Perich, M., Lajoie, G., & Dyer, E. (2024). A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, *36*.

Bashiri, M., Walker, E., Lurz, K.-K., Jagadish, A., Muhammad, T., Ding, Z., Ding, Z., Tolias, A., & Sinz, F. (2021). A flow-based latent state generative model of neural population responses to natural images. *Advances in Neural Information Processing Systems*, *34*.

Batty, E., Merel, J., Brackbill, N., Heitman, A., Sher, A., Litke, A., Chichilnisky, E., & Paninski, L. (2017). Multilayer recurrent network models of primate retinal ganglion cell responses. In *International Conference on Learning Representations*.

Burg, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., & Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, *17*(6), e1009028.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, *15*(4), e1006897.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.*, *10*(12), e1003963.

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 11030–11039).

Cichy, R. M., Dwivedi, K., Lahner, B., Lascelles, A., Iamshchinina, P., Graumann, M., Andonian, A., Murty, N. A. R., Kay, K., Roig, G., & Oliva, A. (2021). The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *arXiv*.

Cichy, R. M., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., Mohsenzadeh, Y., Ramakrishnan, K., & Oliva, A. (2019). The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv*.

Cowley, B., & Pillow, J. (2020). High-contrast "gaudy" images improve the training of deep neural network models of visual cortex. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.) *Advances in Neural Information Processing Systems 33*, (pp. 21591–21603). Curran Associates, Inc.

Dao, T. (2023). Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

de Vries, S. E. J., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., Feng, D., Cain, N., Ledochowitsch, P., Millman, D., Roll, K., Garrett, M., Keenan, T., Kuan, L., Mihalas, S., Olsen, S., Thompson, C., Wakeman, W., Waters, J., Williams, D., Barber, C., Berbesque, N., Blanchard, B., Bowles, N., Caldejon, S. D., Casal, L., Cho, A., Cross, S., Dang, C., Dolbeare, T., Edwards, M., Galbraith, J., Gaudreault, N., Gilbert, T. L., Griffin, F., Hargrave, P., Howard, R., Huang, L., Jewell, S., Keller, N., Knoblich, U., Larkin, J. D., Larsen, R., Lau, C., Lee, E., Lee, F., Leon, A., Li, L., Long, F., Luviano, J., Mace, K., Nguyen, T., Perkins, J., Robertson, M., Seid, S., Shea-Brown, E., Shi, J., Sjoquist, N., Slaughterbeck, C., Sullivan, D., Valenza, R., White, C., Williford, A., Witten, D. M., Zhuang, J., Zeng, H., Farrell, C., Ng, L., Bernard, A., Phillips, J. W., Reid, R. C., & Koch, C. (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat. Neurosci.*, *23*(1), 138–151.

Dean, J., Patterson, D., & Young, C. (2018). A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro*, *38*(2), 21–29.

Ding, Z., Tran, D. T., Ponder, K., Cobos, E., Ding, Z., Fahey, P. G., Wang, E., Muhammad, T., Fu, J., Cadena, S. A., et al. (2023). Bipartite invariance in mouse primary visual cortex. *bioRxiv*.
URL https://www.biorxiv.org/content/10.1101/2023.03.15.532836v1

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nat. Rev. Neurosci.*, *24*(7), 431–450.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
URL https://openreview.net/forum?id=YicbFdNTTy

Ecker, A. S., Sinz, F. H., Froudarakis, E., Fahey, P. G., Cadena, S. A., Walker, E. Y., Cobos, E., Reimer, J., Tolias, A. S., & Bethge, M. (2018). A rotation-equivariant convolutional neural network model of primary visual cortex. *arXiv*.

Elfwing, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, *107*, 3–11.

Franke, K., Willeke, K. F., Ponder, K., Galdamez, M., Zhou, N., Muhammad, T., Patel, S., Froudarakis, E., Reimer, J., Sinz, F. H., & Tolias, A. S. (2022). State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, *610*(7930), 128–134.
URL https://doi.org/10.1038/s41586-022-05270-3

Fu, J., Shrinivasan, S., Ponder, K., Muhammad, T., Ding, Z., Wang, E., Ding, Z., Tran, D. T., Fahey, P. G., Papadopoulos, S., Patel, S., Reimer, J., Ecker, A. S., Pitkow, X., Haefner, R. M., Sinz, F. H., Franke, K., & Tolias, A. S. (2023). Pattern completion and disruption characterize contextual modulation in mouse visual cortex. *bioRxiv*.
URL https://www.biorxiv.org/content/early/2023/03/14/2023.03.13.532473

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193–202.

 https://doi.org/10.52202/079017-3777

George, D., & Hawkins, J. (2005). A hierarchical bayesian model of invariant pattern recognition in the visual cortex. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 3, (pp. 1812–1817). IEEE.

Gifford, A. T., Lahner, B., Saba-Sadiya, S., Vilas, M. G., Lascelles, A., Oliva, A., Kay, K., Roig, G., & Cichy, R. M. (2023). The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*.

Gilbert, C. D., & Wiesel, T. N. (1983). Clustered intrinsic connections in cat visual cortex. *Journal of Neuroscience*, *3*(5), 1116–1133.

Gilbert, C. D., & Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, *9*(7), 2432–2442.

Heeger, D. J. (1992a). Half-squaring in responses of cat striate cells. *Vis. Neurosci.*, *9*(5), 427–443.

Heeger, D. J. (1992b). Normalization of cell responses in cat striate cortex. *Vis. Neurosci.*, *9*(2), 181–197.

Hoefling, L., Szatko, K. P., Behrens, C., Qiu, Y., Klindt, D. A., Jessen, Z., Schwartz, G. S., Bethge, M., Berens, P., Franke, K., et al. (2022). A chromatic feature detector in the retina signals visual context changes. *bioRxiv*, (pp. 2022–11).
URL https://www.biorxiv.org/content/10.1101/2022.11.30.518492.abstract

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 7132–7141).

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, (pp. 646–661). Springer.

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, *58*(6), 1187–1211.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-Scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1725–1732).

Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. *Journal of vision*, *19*(4), 29–29.

Klindt, D. A., Ecker, A. S., Euler, T., & Bethge, M. (2017). Neural system identification for large populations separating "what" and "where". In *Advances in Neural Information Processing Systems*, (pp. 4–6).

Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*.

Li, B. M., Cornacchia, I. M., Rochefort, N., & Onken, A. (2023). V1t: large-scale mouse v1 response prediction using a vision transformer. *Transactions on Machine Learning Research*.
URL https://openreview.net/forum?id=qHZs2p4ZD4

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, *18*(1), 6765–6816.

Liu, J. K., Schreyer, H. M., Onken, A., Rozenblit, F., Khani, M. H., Krishnamoorthy, V., Panzeri, S., & Gollisch, T. (2017). Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nature communications*, *8*(1), 149.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 3202–3211).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lurz, K.-K., Bashiri, M., Willeke, K., Jagadish, A. K., Wang, E., Walker, E. Y., Cadena, S. A., Muhammad, T., Cobos, E., Tolias, A. S., Ecker, A. S., & Sinz, F. H. (2021). Generalization in data-driven models of primary visual cortex. In *Proceedings of the International Conference for Learning Representations (ICLR)*, (p. 2020.10.05.326256).

Marques, T., Nguyen, J., Fioreze, G., & Petreanu, L. (2018). The functional organization of cortical feedback inputs to primary visual cortex. *Nature neuroscience*, *21*(5), 757–764.

McIntosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S., & Baccus, S. A. (2016). Deep learning models of the retinal response to natural scenes. *Adv. Neural Inf. Process. Syst.*, *29*(Nips), 1369–1377.

MICrONS Consortium, Alexander Bae, J., Baptiste, M., Bodor, A. L., Brittain, D., Buchanan, J., Bumbarger, D. J., Castro, M. A., Celii, B., Cobos, E., Collman, F., da Costa, N. M., Dorkenwald, S., Elabbady, L., Fahey, P. G., Fliss, T., Froudakis, E., Gager, J., Gamlin, C., Halageri, A., Hebditch, J., Jia, Z., Jordan, C., Kapner, D., Kemnitz, N., Kinn, S., Koolman, S., Kuehner, K., Lee, K., Li, K., Lu, R., Macrina, T., Mahalingam, G., McReynolds, S., Miranda, E., Mitchell, E., Mondal, S. S., Moore, M., Mu, S., Muhammad, T., Nehoran, B., Ogedengbe, O., Papadopoulos, C., Papadopoulos, S., Patel, S., Pitkow, X., Popovych, S., Ramos, A., Clay Reid, R., Reimer, J., Schneider-Mizell, C. M., Sebastian Seung, H., Silverman, B., Silversmith, W., Sterling, A., Sinz, F. H., Smith, C. L., Suckow, S., Tan, Z. H., Tolias, A. S., Torres, R., Turner, N. L., Walker, E. Y., Wang, T., Williams, G., Williams, S., Willie, K., Willie, R., Wong, W., Wu, J., Xu, C., Yang, R., Yatsenko, D., Ye, F., Yin, W., & Yu, S.-C. (2021). Functional connectomics spanning multiple areas of mouse visual cortex. *bioRxiv*, (p. 2021.07.28.454025).

Morrone, M. C., Tosetti, M., Montanaro, D., Fiorentini, A., Cioni, G., & Burr, D. C. (2000). A cortical area that responds specifically to optic flow, revealed by fMRI. *Nat. Neurosci.*, *3*(12), 1322–1328.

Niell, C. M., & Stryker, M. P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, *65*(4), 472–479.

Pavao, A., Guyon, I., Letournel, A.-C., Baró, X., Escalante, H., Escalera, S., Thomas, T., & Xu, Z. (2022). *CodaLab Competitions: An open source platform to organize scientific challenges*. Ph.D. thesis, Université Paris-Saclay, FRA.

Pei, F., Ye, J., Zoltowski, D., Wu, A., Chowdhury, R. H., Sohn, H., O'Doherty, J. E., Shenoy, K. V., Kaufman, M. T., Churchland, M., et al. (2021). Neural latents benchmark'21: evaluating latent variable models of neural population activity. *arXiv preprint arXiv:2109.04463*.

Perrone, J. A., & Liston, D. B. (2015). Redundancy reduction explains the expansion of visual direction space around the cardinal axes. *Vision Research*, *111*, 31–42.

Petkov, N., & Subramanian, E. (2007). Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal gabor filters with surround inhibition. *Biol. Cybern.*, *97*(5-6), 423–439.

Pogoncheff, G., Granley, J., & Beyeler, M. (2023). Explaining v1 properties with a biologically constrained deep learning architecture. *Advances in Neural Information Processing Systems*, *36*, 13908–13930.

Qiu, Y., Klindt, D. A., Szatko, K. P., Gonschorek, D., Hoefling, L., Schubert, T., Busse, L., Bethge, M., & Euler, T. (2023). Efficient coding of natural scenes improves neural system identification. *PLOS Computational Biology*, *19*(4), e1011037.

Reimer, J., Froudarakis, E., Cadwell, C. R., Yatsenko, D., Denfield, G. H., & Tolias, A. S. (2014). Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron*, *84*(2), 355–362.

 https://doi.org/10.52202/079017-3777

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, *115*(3), 211–252.

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, *46*(6), 945–956.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4510–4520).

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, (p. 407007).

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423.

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., Groblewski, P. A., Ahmed, R., Arkhipov, A., Bernard, A., Billeh, Y. N., Brown, D., Buice, M. A., Cain, N., Caldejon, S., Casal, L., Cho, A., Chvilicek, M., Cox, T. C., Dai, K., Denman, D. J., de Vries, S. E. J., Dietzman, R., Esposito, L., Farrell, C., Feng, D., Galbraith, J., Garrett, M., Gelfand, E. C., Hancock, N., Harris, J. A., Howard, R., Hu, B., Hytnen, R., Iyer, R., Jessett, E., Johnson, K., Kato, I., Kiggins, J., Lambert, S., Lecoq, J., Ledochowitsch, P., Lee, J. H., Leon, A., Li, Y., Liang, E., Long, F., Mace, K., Melchior, J., Millman, D., Mollenkopf, T., Nayan, C., Ng, L., Ngo, K., Nguyen, T., Nicovich, P. R., North, K., Ocker, G. K., Ollerenshaw, D., Oliver, M., Pachitariu, M., Perkins, J., Reding, M., Reid, D., Robertson, M., Ronellenfitch, K., Seid, S., Slaughterbeck, C., Stoecklin, M., Sullivan, D., Sutton, B., Swapp, J., Thompson, C., Turner, K., Wakeman, W., Whitesell, J. D., Williams, D., Williford, A., Young, R., Zeng, H., Naylor, S., Phillips, J. W., Reid, R. C., Mihalas, S., Olsen, S. R., & Koch, C. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, *592*(7852), 86–92.

Simoncelli, E. P., Paninski, L., Pillow, J., Schwartz, O., et al. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, *3*(327-338), 1.

Sinz, F., Ecker, A. S., Fahey, P., Walker, E., Cobos, E., Froudarakis, E., Yatsenko, D., Pitkow, Z., Reimer, J., & Tolias, A. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in neural information processing systems*, *31*.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*(6), 967–979.
URL https://doi.org/10.1016/j.neuron.2019.08.034

Sofroniew, N. J., Flickinger, D., King, J., & Svoboda, K. (2016). A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *elife*, *5*, e14472.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, (pp. 6105–6114). PMLR.

Touryan, J., Felsen, G., & Dan, Y. (2005). Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, *45*(5), 781–791.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, (pp. 6450–6459).

Turishcheva, P., Fahey, P. G., Hansel, L., Froebe, R., Ponder, K., Vystrčilová, M., Willeke, K. F., Bashiri, M., Wang, E., Ding, Z., et al. (2023). The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. *ArXiv*.

Ustyuzhaninov, I., Burg, M. F., Cadena, S. A., Fu, J., Muhammad, T., Ponder, K., Froudarakis, E., Ding, Z., Bethge, M., Tolias, A. S., & Ecker, A. S. (2022). Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex.
URL https://doi.org/10.1101/2022.02.10.479884

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vintch, B., Movshon, J. A., & Simoncelli, E. P. (2015). A convolutional subunit model for neuronal responses in macaque v1. *Journal of Neuroscience*, *35*(44), 14829–14841.

Vystrčilová, M., Sridhar, S., Burg, M. F., Gollisch, T., & Ecker, A. S. (2024). Convolutional neural network models of the primate retina reveal adaptation to natural stimulus statistics. *bioRxiv*.
URL https://www.biorxiv.org/content/early/2024/03/09/2024.03.06.583740

Walker, E. Y., Cotton, R. J., Ma, W. J., & Tolias, A. S. (2020). A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, *23*(1), 122–129.

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.*, *22*(12), 2060–2065.

Wang, B. (2021). Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, E. Y., Fahey, P. G., Ponder, K., Ding, Z., Chang, A., Muhammad, T., Patel, S., Ding, Z., Tran, D., Fu, J., Papadopoulos, S., Franke, K., Ecker, A. S., Reimer, J., Pitkow, X., Sinz, F. H., & Tolias, A. S. (2023). Towards a foundation model of the mouse visual cortex. *bioRxiv*.
URL https://www.biorxiv.org/content/early/2023/03/24/2023.03.21.533548

Willeke, K. F., Fahey, P. G., Bashiri, M., Pede, L., Burg, M. F., Blessing, C., Cadena, S. A., Ding, Z., Lurz, K.-K., Ponder, K., Muhammad, T., Patel, S. S., Ecker, A. S., Tolias, A. S., & Sinz, F. H. (2022). The sensorium competition on predicting large-scale mouse primary visual cortex activity. *arXiv*.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, *29*, 477–505.

Wu, N., Valera, I., Ecker, A., Euler, T., & Qiu, Y. (2023). Bayesian neural system identification with response variability. *arXiv preprint arXiv:2308.05990*.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.
URL https://doi.org/10.1073/pnas.1403112111

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, (pp. 6023–6032).

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018a). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 6848–6856).

Zhang, Y., Lee, T.-S. T. S., Li, M., Liu, F., Tang, S., Sing, T., Ming, L., Fang, L., Shiming, L., Lee, T.-S. T. S., Li, M., Liu, F., & Tang, S. (2018b). Convolutional neural network models of V1 responses to complex patterns. *J. Comput. Neurosci.*, (pp. 1–22).

Zheng, Y., Jia, S., Yu, Z., Liu, J. K., & Huang, T. (2021). Unraveling neural coding of dynamic natural visual scenes via convolutional recurrent neural networks. *Patterns*, *2*.

 https://doi.org/10.52202/079017-3777

A Appendix

A.1 First place – DwiseNeuro

Analysis of Improvements. All of the score numbers are in for the main track during the competition live phase. An early model with depth-wise 3D convolution blocks achieved a score of ≈0.19. Implementing techniques from the core section, tuning hyperparameters, and training on ten mice instead of five data boosted the score to 0.25. Removing normalization improved the score to 0.27. The cortex and CutMix (Yun et al., 2019) increased the score to 0.276. Then, the β value of Softplus was tuned, resulting in a score of 0.294. Lastly, adjusting drop rate and batch size parameters helped to achieve a score of 0.3. The ensemble of the basic and distillation A.1 training stages achieved a single-trial correlation of 0.2913. This is just slightly better than the basic training.

- Learning rate warmup for the first three epochs from 0 to 2.4e-03
- cosine annealing last 18 epochs to 2.4e-05
- Batch size 32, one training epoch comprises 72000 samples
- Optimizer AdamW with weight decay 0.05
- Poisson loss
- Model EMA with decay 0.999
- CutMix with alpha 1.0 and usage probability 0.5
- The sampling of different mice in the batch is random by uniform distribution

A.2 Second place – The Runner-up Solution *Dynamic-V1FM*

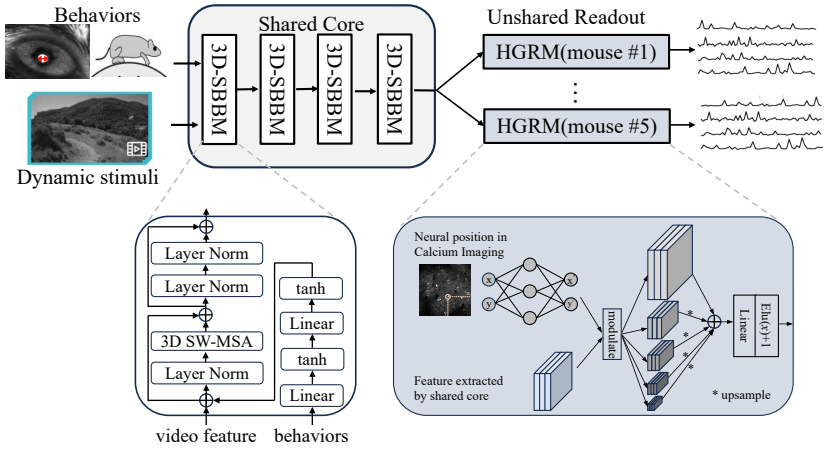


Figure 4: The overall architecture of Dynamic-V1FM. The core module consists of four 3D-SBBMs that process video and behavioral information, as detailed in the lower left. The unshared readout module includes five levels of features before linear readout in the lower right.

A.2.1 Experiments

Training Details. We trained Dynamic-V1FM using the training set of ten mice data provided by the competition, and tested it with only five mice data required for submission. Note that we did not employ any pre-training strategy and directly performed the evaluations required by the competition after training. During training, we used truncated normal initialization for the core module and the same initialization strategy for the readout module as Lurz et al. (2021). The whole model was optimized by AdamW optimizer (Loshchilov & Hutter, 2017) with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight_decay $= 0.05$, and a batch size of 32. Each batch contained 30 frames of randomly sampled data. The peak learning rate was $1e^{-3}$, linearly warmed up with ratio $\frac{1}{3}$ for the first 600 iterations, then kept constant for the first 80 epochs and decreased to 10^{-6} in the last 100 epochs with a cosine strategy. All the models used in the ensemble strategy shared the same training setting.

Experimental Results. On the live-test evaluation, the improvement of the core module, replacing 3D convolution with 3D swin transformer, resulted in an R^2 improvement of 0.045 (from 0.188 to 0.233). Enhancements in the readout module, replacing Gaussian readout to Hierarchical Gaussian readout, further improved the model by 0.018 (from 0.233 to 0.251). The final ensemble strategy yielded an overall prediction score of 0.276.

Parameter Search: The participants did a limited hyperparameter search; they manually adjusted only the learning rate (1e-4,5e-4,1e-3,2e-3,5e-3) and the number of channels in the Core module (32,64,128,160,224). As for the batch size, the GPU utilization was maximized, therefore, batch size was 32 consistently throughout the experiments. All other hyperparameters were left at their default values without further optimization.

A.2.2 Discussions

We shall provide some thoughts on the V1FM design. Using combined data sets from multiple mice and a *shared core* module for training is an efficient approach, although the subject-specific readout module strategy increases the difficulty of training the core module. This design could be viewed as a stronger regularization that may weaken the performance of the whole model. This problem might be alleviated by designing a new shared readout module that also relies on subject-specific information, such as mice identities and behavioral data. Specifically, we can use a readout module with dynamic weights (Chen et al., 2020) which is adjusted by mice identities and pupil size.

A.3 Third place – ViV1T

Table 3: ViV1T core hyperparameter search space and their final settings. We performed Hyperband Bayesian optimization (Li et al., 2017) with 20 iterations to find the setting that yield the best single trial correlation in the validation set. The resulting ViV1T model contains 12M trainable parameters, about 13% more than the factorized baseline.

HYPERPARAMETER	SEARCH SPACE	FINAL VALUE
CORE		
EMBEDDING DIM.	UNIFORM, MIN: 8, MAX: 512, STEP: 8	112
LEARNING RATE	UNIFORM, MIN: 0.0001, MAX: 0.01	0.0048
PATCH DROPOUT	UNIFORM, MIN: 0, MAX: 0.5	0.1338
DROP PATH	UNIFORM, MIN: 0, MAX: 0.5	0.0505
POS. ENCODING	NONE, LEARNABLE, SINUSOIDAL	LEARNABLE
WEIGHT DECAY	UNIFORM, MIN: 0, MAX: 1	0.1789
BATCH SIZE	UNIFORM, MIN: 1, MAX: 64	6
SPATIAL TRANSFORMER		
NUM. BLOCKS	UNIFORM, MIN:1, MAX: 8, STEP: 1	3
PATCH SIZE	UNIFORM, MIN: 3, MAX: 16, STEP: 1	7
PATCH STRIDE	UNIFORM, MIN: 1, MAX: PATCH SIZE, STEP: 1	2
TEMPORAL TRANSFORMER		
NUM. BLOCKS	UNIFORM, MIN:1, MAX: 8, STEP: 1	5
PATCH SIZE	UNIFORM, MIN: 1, MAX: 50, STEP: 1	25
PATCH STRIDE	UNIFORM, MIN: 1, MAX: PATCH SIZE, STEP: 1	1
MULTI-HEAD ATTENTION (MHA) LAYER		
NUM. HEADS	UNIFORM, MIN: 1, MAX: 16, STEP: 1	11
HEAD DIM.	UNIFORM, MIN: 8, MAX: 512, STEP: 8	48
MHA DROPOUT	UNIFORM, MIN: 0, MAX: 0.5	0.3580
FEEDFORWARD (FF) LAYER		
FF DIM.	UNIFORM, MIN: 8, MAX: 512, STEP: 8	136
FF ACTIVATION	TANH, SIGMOID, ELU, GELU, SWIGLU	GELU
FF DROPOUT	UNIFORM, MIN: 0, MAX: 0.5	0.0592

https://doi.org/10.52202/079017-3777

HYPERPARAMETER	IMPORTANCE	CORRELATION
batch size	0.374	-0.663
core learning rate	0.099	-0.411
weight decay	0.063	0.259
drop path	0.041	-0.353
embedding dim.	0.039	0.040
FF dropout	0.031	-0.358
patch dropout	0.030	-0.339
FF dim.	0.026	0.251
temporal patch size	0.026	0.104
learning rate	0.024	-0.265

Table 4: **The top 10 most important hyperparameters estimated in the Bayesian hyperparameter search for the ViV1T solution** (3rd place). The importance score is the feature importance values of a random forest model trained on the hyperparameters as inputs and the hyperparameter search objective (single trial correlation on the validation set) as the target output. The correlation value is the linear correlation between the hyperparameter and search objective.

Baseline	Core	Layers	Channels	Input Spatial Kernels	Spatial Kernels
GRU	Rotation-equivariant	4	8	9×9	7×7
3D Factorized	3D factorized	3	32, 64, 128	11×11	5×5

Table 5: **Core parameters for the baseline architectures**. Compared to the GRU baseline, the amount of channels in the core was increased sequentially.

A.4 Baseline architectures parameters

The **GRU baseline** used rotation-equivariant core from Ecker et al. (2018) with 8 rotations, resulting in 64 channels totally (8 channels $\times$ 8 rotations = 64). Inspired by Sinz et al. (2019) we used the GRU module after the core. It had 64 channels, and both input and recurrent kernels were 9×9.

For the **3D Factorized baseline**, we used the core inspired by Hoefling et al. (2022); Vystrčilová et al. (2024). The temporal kernels were 11×1 in the 1st layer and 5×1 afterwards, same as the spatial ones (Tab. 5).

The Ensembled baseline cores were same as for the 3D Factorized baseline.

A.5 Stability analysis

	Main track	
Seed	single-trial ρ_{st} ↑	average ρ_{ta} ↑
8	0.1932	0.3650
16	0.1642	0.3210
42	0.1887	0.3569
64	0.1780	0.3380
128	0.1839	0.3479
512	0.1799	0.3402
1024	0.1865	0.3528
2048	0.1672	0.3178
4096	0.1734	0.3305
16384	0.1880	0.3571
32768	0.1933	0.3661
131072	0.1852	0.3513
262144	0.1839	0.3488
1048576	0.1943	0.3674
mean	0.1828	0.3472
std	0.0094	0.0159

Table 6: We used seeds 8, 16, 42, 64, 128, 512, 1024, 2048, 4096, 16384, 32768, 131072, 262144, 104857 to ensemble the factorized benchmark. Here we provide the individual performance of the models on the final test set to analyse how much performance depended on the seed.

A.6 Discussion on what is crucial to create the best performing model

Indeed hyperparameter tuning is an important part of succeeding in a competition. Unfortunately, it is really hard to compare the degree of hyperparameter tuning across different models, especially when it comes to hyperparameters such as number of layers or layers width because the solutions used fundamentally different architectures (CNN vs transformers). Two of the winning teams varied the batch size, dropout and weight decay, therefore, we believe this is not the main factor determining the winner.

Usage of data augmentation is still very limited in the neuronal predictive modeling community because it is not clear how augmenting the image should change the neurons responses. In fact, this is the first time we saw data augmentation (CutMix) used for this task and we are happy to report it, though its improvement was fairly limited according to appendix A1 (+0.006). The biological meaning of CutMix also requires further investigation.

Tuning the nonlinearity is an interesting strategy. Tuning the softplus seemed to be quite helpful (+0.018). This result is not entirely unexpected: Cadena et al. (2019) also *"found that optimizing the final nonlinearity, $f(x)$, of each neuron was important for optimal performance of the data-driven CNN"*. However, this adds parameters, therefore, it is partially an architectural change as well.

https://doi.org/10.52202/079017-3777

A.7 Existing datasets comparison

SOURCE	GOAL	RECORDING	SPECIES	BRAIN AREA	NEURONS	STIMULI / TASK	TEST SET
Sensorium 2022	Model comparision for static stimuli responses in mice V1	Calcium imaging	7 mice	V1	>28,000	Natural images	+
Allen Brain Observatory	Study how mouse visual cortex works	Calcium imaging	243 mice	V1, LM, AL, PM, AM, RL, layers 2-6	∼60,000	Drifting/Static gratings Natural images/movies; Spontaneuos activity; Locally sparse noise;	–
MICrONS	Study how mouse visual cortex works	Calcium imaging	1 mouse	V1, RL, AL, LM; layers 1-6	∼75,000	Natural Movies; Local /Global Directional Parametric Stimulus	–
Berkeley Neural Prediction Challenge	Compare neural predictive models for different systems	Electro-physio-logy	Macaque songbird ferret	primary visual & auditory cortex, field L	208; only 12 for V1 movies	Natural movies/images; birdsong or 3-sec natural vocalizations	+
Brain-Score	Evaluate how well task-pretrained com-puter vision models match the neural representations	Electro-physio-logy & fMRI	Human & non-hum. primates	V1, V2, V4, IT	2018: 256 neurons. 2022 & 2024 included fMRI	2018 & 2024: natural 2022 had a few videos but images were the main focus	2018 + 2022 + 2024 -
Neural Latents 21	Predict held-out neurons' activity from other neurons	Electro-physio-logy	monkeys	dorsal premotor, primary motor cortex Brodmann's area 2, dorsomedial frontal cort.	431	motor tasks: maze; random target; bump task; ready-set-go tasks	+
Algonauts 23	Model comparison for fMRI responses (used The Natural Scenes Dataset (NSD))	fMRI	8 Humans	whole brain	–	Natural scenes	+
Sensorium 2023 (Ours)	Models comparision for dynamic stimuli responses in mice V1	Calcium imaging	10 mice	V1	78,853	Natural movies/images; Global Directional Parametric Stimulus; Random/Gaussian Dots; Drifting Gabor	+

Table 7: **Comparing existing datasets.** None of the previously existing datasets could be used for Sensorium 2023 competition as they do not have enough neurons, image different species or brain areas, do not have single-neuron resolution or do not have held-out test set. Data source can be a competition, agregating several datasets or a standalone dataset.
PS: Brain Score '24 focused on creating benchmarks, while 18 and 22 compared models.

https://doi.org/10.52202/079017-3777 118926

A.8 Supplementary materials

A.8.1 Dataset documentation and intended uses

Dataset documentation is available at https://gin.g-node.org/pollytur/sensorium_2023_dataset (dataset stucture) and in the whitepaper (Turishcheva et al., 2023) (data collection methodology). Intended usage examples (loading of the data and models training) are available here: https://github.com/ecker-lab/sensorium_2023/tree/main/notebooks.

A.8.2 URL for data download

Five competition mice: https://gin.g-node.org/pollytur/sensorium_2023_dataset
Five mice with ood responses https://gin.g-node.org/pollytur/sensorium_2023_data/src/798ba8ad041d8f0f0ce879af396d52c7238c2730.

A.8.3 Croissant url

As the croissant library currently does not support the Video and List data types (https://github.com/mlcommons/croissant/issues/690), we generated the high-level meta file using kaggle interface: https://github.com/ecker-lab/sensorium_2023/blob/croissant_file/sensorium-2023-metadata.json

A.8.4 Author statement

Author bear all responsibility in case of violation of rights. Both data and code are available under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

A.8.5 Hosting, licensing, and maintenance plan

Following SENSORIUM 2022, data is hosted at https://gin.g-node.org, which is a publicly available platform, where data can be downloaded both via GUI or command line. The code is hosted in a public repository via https://github.com. The data does not need maintenance. Both data and code are available under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. In case of any problems with the data hosting webpage, the authors have local copies of data and would re-release it.

118927 https://doi.org/10.52202/079017-3777

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   [Yes] Yes, we introduce a competition to predict single neuron activity in visual cortex in response to dynamic stimuli. The competition is accompanied with a unique dataset and then stay as a benchmark after the competition end. We report the improvements achieved during the competition.
   (b) Did you describe the limitations of your work?
   [Yes] In Sec.7 we discuss that openly available training infrastructure might not be enough for our competition and in Sec.8 we provide the ways to extend our benchmark.
   (c) Did you discuss any potential negative societal impacts of your work?
   [No] We do not think there are potential negative societal impacts of our work. The benchmarks have shown to improve the machine learning field and the model needed for the sake of this competition/benchmark are small and do not require a lot of resources in comparision with modern LLM, picture or video generation models, etc.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   [Yes] We read and followed NeurIPS Code of Ethics. Moreover, as our dataset involved animal experiment, all procedures were approved by the Institutional Animal Care and Use Committee of Baylor College of Medicine.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?
   [N/A] Paper does not provide theoretical results.
   (b) Did you include complete proofs of all theoretical results?
   [N/A] Paper does not provide theoretical results.
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   [Yes] We include URL for the data access and for the benchmark code and participants also provided URL for their code (Sections 3 and 6).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   [Yes] The data split is fixed with the dataset. The training procedure is described in Sec.5. The winners training procedures are decribed in Sec.6 if different and in appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   [Yes] We do not provide error bars but we do provide the analysis of seed impact in appendix, section "Stability analysis" (Tab. 6)
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
   [Yes] . See table 2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   [Yes] We used Codalab Competition to self-host our competition webpage and the Codalab Competition is mentioned and cited in 'Sensorium Competition Overview' section (section 2).

https://doi.org/10.52202/079017-3777 118928

(b) Did you mention the license of the assets?
[Yes] . Yes, Apache License 2.0 is mentioned at the same place as Codalab Competition is cited.

(c) Did you include any new assets either in the supplemental material or as a URL?
[Yes] We include URL for the data access and for the benchmark code and participants also provided URL for their code (Sections 3 and 6).

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
[N/A] We do not use any data from people.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
[N/A] The dataset we provide does not contain any personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable?
[N/A] We did not used crowdsourcing or research with human subjects

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?
[N/A] We did not used crowdsourcing or research with human subjects

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
[N/A] We did not used crowdsourcing or research with human subjects

118929 https://doi.org/10.52202/079017-3777